DOCUMENT RESUME

ED 474 869 TM 034 811

AUTHOR	Koretz, Daniel M.; Barton, Karen
TITLE	Assessing Students with Disabilities: Issues and Evidence. CSE Technical Report.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.; National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	CSE-TR-587
PUB DATE	2003-01-00
NOTE	37p.
CONTRACT	R305B960002-01
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Academic Accommodations (Disabilities); *Disabilities; Elementary Secondary Education; *Large Scale Assessment; State Programs; Testing Programs

ABSTRACT

Until recently, many students with disabilities were excluded from large-scale assessments, such as those mandated by states. Recent federal and state policy initiatives, including the most recent reauthorization of the Individuals with Disabilities Education Act, require that the large majority of students with disabilities be included in statewide assessments used in accountability systems. Most observers agree that educational outcomes for students with disabilities were inadequate before the new policies were implemented; however, the research undergirding the new policies is limited. The reforms have spurred a rapid increase in relevant research, but more and improved research is needed. This paper reviews the status of research on some classification and assessment issues that are central to the new reforms and recommends directions for future research. Better descriptive information is needed about the target populations, and much research is needed on the assessment of disabilities, focusing on validity issues. (Contains 3 tables and 68 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

ED 474 869

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

K. Hurst

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Assessing Students With Disabilities: Issues and Evidence

CSE Technical Report 587

Daniel M. Koretz
CRESST/Harvard Graduate School of Education
Karen Barton
CTB/McGraw-Hill

TM034811

UCLA Center for the Study of Evaluation

In Collaboration With:

University of Colorado at Boulder • Stanford University • The RAND Corporation
University of Southern California • Educational Testing Service
University of Pittsburgh • University of Cambridge

BEST COPY AVAILABLE

Assessing Students With Disabilities:
Issues and Evidence

CSE Technical Report 587

Daniel M. Koretz
CRESST/Harvard Graduate School of Education
Karen Barton
CTB/McGraw-Hill

January 2003

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.1 Comparative Analyses of Current Assessment and Accountability Systems/Strand 3
Daniel Koretz, Project Director, CRESST/Harvard Graduate School of Education

Copyright © 2003 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

ASSESSING STUDENTS WITH DISABILITIES: ISSUES AND EVIDENCE

Daniel M. Koretz
CRESST/Harvard Graduate School of Education

Karen Barton
CTB/McGraw-Hill

Abstract

Until recently, many students with disabilities were excluded from large-scale assessments, such as those mandated by states. Recent federal and state policy initiatives, including the most recent reauthorization of IDEA, require that the large majority of students with disabilities be included in the statewide assessments used in accountability systems. Most observers agree that educational outcomes for students with disabilities were inadequate before the new policies were implemented; however, the research undergirding the new policies is limited. The reforms have spurred a rapid increase in relevant research, but more and improved research is needed. This paper reviews the status of research on some classification and assessment issues that are central to the new reforms and recommends directions for future research.

Until very recently, many students with disabilities were routinely excluded from large-scale assessments. Guidelines pertaining to the exclusion of students with disabilities from statewide assessments differed from one state to another, and the estimated rate of participation of students with disabilities varied markedly across states and was often low (Erickson, Thurlow, & Thor, 1995; McGrew, Thurlow, Shriner, & Spiegel, 1992; Shriner & Thurlow, 1992). In addition, decisions about whether to include students with disabilities were often made by local school personnel, such as the team responsible for students' Individualized Education Programs, or IEPs (Erickson & Thurlow, 1996), and this introduced additional variation in patterns of inclusion. In many cases, educators faced incentives to exclude from assessments students with disabilities who may score poorly.

Recent reforms at the national and state level, however, have attempted to increase the inclusion of students with disabilities in large-scale assessments, as part of a nationwide movement to include all children in standards-based reforms. For example, the 1997 amendments to the Individuals with Disabilities Education Act ([IDEA] 1997) stipulated that a state is eligible for assistance under Part B, the main

federal grant program for the education of students with disabilities, only if it meets the following requirements:

- The State has established goals for the performance of children with disabilities in the State that are consistent, to the maximum extent appropriate, with other goals and standards for children established by the State;
- [The State] has established performance indicators the State will use to assess progress toward achieving those goals that, at a minimum, address the performance of children with disabilities on assessments, drop-out rates, and graduation rates;
- Children with disabilities are included in general State and district-wide assessment programs, with appropriate accommodations, where necessary;
- [The State] develops and, beginning not later than July 1, 2000, conducts . . . alternate assessments . . . for those children [with disabilities] who cannot participate in State and district-wide assessment programs; [and]
- The State educational agency . . . reports to the public . . . the number of children with disabilities participating in regular assessments . . . [and] the performance of those children on regular assessments. (20 U.S.C. 1400)

Some state policymakers began steps to improve inclusion in assessments even before the passage of the IDEA Amendments of 1997 required it. Both Kentucky and Maryland, for example, were building systems similar to those eventually specified by IDEA, requiring that most students with disabilities be tested with the regular state assessment and providing an alternate assessment for the small minority of students with disabilities deemed unable to participate in the regular assessment, several years before the 1997 reauthorization.

These efforts to increase the inclusion of students with disabilities in large-scale assessments are motivated by several goals. It is hoped that inclusion will provide better information not only about the performance of students with disabilities, but also about the aggregate performance of their schools.[1] Perhaps more important, proponents of these changes hope that including these students in assessments—especially, the large-scale assessments tied to accountability in standards-based reforms—will make schools more accountable for, and thus more attentive to, the academic performance of students with disabilities.

[1]For brevity, we use the term "inclusion" to refer to inclusion in large-scale assessments, not to inclusion in general-education settings.

Several years ago, a National Research Council study committee noted that these efforts to increase the participation of students with disabilities in large-scale assessments were hindered by a lack of experience and research-based information (McDonnell, McLaughlin, & Morison, 1997). Despite a growing amount of relevant research in the past few years, this caution still holds true. For example, there is only limited systematic information about the use of testing accommodations for elementary and secondary students with disabilities and even less about the effects of accommodations on the validity of scores. Nor is there systematic evidence about the effects of inclusion on the opportunities afforded to students with disabilities or on their educational achievement and attainment.

In this paper, we outline major issues raised by the inclusion of students with disabilities in large-scale assessments and summarize some of the pertinent research. We then describe an agenda for future research and discuss implications for policy and practice.

Major Issues in Assessing Students With Disabilities

Inclusion raises four particularly important sets of issues:

- issues of identification and classification;
- questions about the appropriate use of accommodations;
- the problem of disabilities that are related to measured constructs; and
- issues pertaining to test design.

Identification and Classification

We use "identification" to refer to the determination that a student has a recognized disability. Although there are several criteria one could use in making this decision, identification usually refers to the decision that a student has a disability under the terms of either the Individuals with Disabilities Education Act (IDEA) or Section 504 of the Rehabilitation Act of 1973. The large majority of identified students is identified under IDEA. In contrast, we use "classification" to refer to the categorization of an identified student's specific disability or disabilities.

Identification is highly inconsistent, raising concerns about over-, under-, and mis-identification. While some of this inconsistency occurs at the level of teachers and schools (Clarizio & Phillips, 1992; McDonnell et al., 1997; Shepard, 1989; Ysseldyke & Algozzine, 1982), there are striking differences even when

identification rates are aggregated to the level of states. Nationally (in the 50 states and the District of Columbia), 11.2% of students between the ages of 6 and 17 years were served under Part B (the main state grant program) of IDEA in the 1999-2000 school year (Table 1). This percentage, however, ranged from a low of 9.1% to a high of almost 16%—a pattern that has been quite consistent for years.

The reported prevalence of specific disability categories shows even greater inconsistency from state to state. Using the categories required for federal reporting, the highest prevalence is of specific learning disabilities. Students with specific learning disabilities constitute more than half of all students served under IDEA Part B. (Part B is Assistance for Education of All Children with Disabilities, the core section of IDEA that provides most of the IDEA funding to states.) The percentage of students identified as learning disabled, however, varies more than threefold among the states, from a low of 3.0% to a high of 9.1% (Table 1). The discrepancies are even larger in the case of less common disabilities. For example, the reported prevalence of mental retardation among students ages 6-17 years varies tenfold, from 0.3% in New Jersey to 3.0% in West Virginia.

These large differences among states in identification and classification rates appear to reflect the influence of differences in policy. There is no reason to expect true prevalence rates to vary greatly from state to state. Variations in prevalence stemming from idiosyncratic decisions by local school personnel would tend to average out when aggregating to the level of entire states.

Such dramatic inconsistencies in identification and classification rates make it difficult to determine how best to assess students with disabilities. Increased

Table 1

Percentage of Students Ages 6-17 years Served Under IDEA, Part B, States and U.S. (50 States and the District of Columbia), 1999-2000

	Lowest state	Highest state	U.S. total
All	9.1 (CO)	15.6 (RI)	11.3
Specific learning disability	3.0 (KY)	9.1 (RI)	5.7
Speech/language	1.0 (IA)	3.9 (WV)	2.3
Mental retardation	0.3 (NJ)	3.0 (WV)	1.1
Emotional disturbance	0.1 (AR)	1.9 (VT)	0.9

Note. From U.S. Department of Education, *To assure the free appropriate public education of all children with disabilities: 23rd Annual report to Congress on the implementation of the Individuals With Disabilities Education Act.* 2001, Table AA10. Available online at http://www.ed.gov/offices/OSERS/OSEP/Products/OSEP2001An1Rpt/Appendix_A_Pt1.pdf

inclusion should be carried out in ways that maximize both the quality of the performance information yielded by the assessment and the net positive effects on the education of the students (i.e., "evidential" and "consequential" validity; Messick, 1989). Success in meeting these goals will hinge on a number of decisions. For example, how many students should be deemed unable to participate in the regular general-education assessment? How might the regular assessment be designed to improve its suitability for students with special needs? Which students should receive special accommodations when they are tested, and what accommodations should they be offered? Making these decisions well requires information on the characteristics and needs of identified students. As explained below, this information is obscured by these large inconsistencies in identification and classification.

The problem of inconsistent classification underscores a standing tension in policies pertaining to students with disabilities—that is, the tension between individualized and centralized decision making. By law, many decisions about individual students with disabilities are made by local personnel—in particular, the members of the team responsible for each student's IEP. These decisions include, for example, the accommodations provided in instruction and assessments. Yet states have retained responsibility for many aspects of the assessment of students with disabilities, such as establishing rules governing exclusion and setting guidelines for the use of accommodations in assessment. In making decisions about students, teachers and others in the school can rely on knowledge about individual students that is inaccessible to policymakers. For this reason, inconsistent classifications may be less problematic for local school personnel in deciding on services for individual children, and indeed many advocates urge educators not to focus on classifications in making such decisions. Local school personnel, however, are unlikely to have some of the information needed to make optimal decisions about assessments; for example, they are unlikely to know the findings of ongoing research on the effects of accommodations. Policymakers may have more access to such information, but they lack detailed knowledge of individual students and are constrained to use information such as classifications in establishing guidelines.

The Use of Accommodations

Tests are often administered to students with disabilities in nonstandard ways. Although the labeling of these departures from standardization has not been consistent, the current edition of the *Standards for Educational and Psychological*

Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) suggests using the term "accommodations" to refer to all of them:

> Here accommodation is used as the general term for any action taken in response to a determination that an individual's disability requires a departure from established testing protocol. Depending on circumstances, such accommodation may include modification of test administration processes or modification of test content. (p. 101)

Actual modifications to the tests may include, for example, testing students with forms normally used for an earlier grade ("out-of-level" or "out-of-grade" testing) or deleting some items from the test. The more common accommodations entail not alterations to the test itself, but rather changes in the presentation or administration of the test or in the student's mode of response. Examples include providing students with additional time; administering the test in a separate location; breaking the testing time into shorter periods with more breaks; reading either directions or actual test items to students; providing the test in a different format, such as Braille or large type; and allowing students to dictate rather than write their responses.

The terminology suggested by the *Standards* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) is a departure from the previously common usage. Before the publication of the current edition of the *Standards*, it was common to use "accommodation" to refer to changes in presentation, response, or setting and to use "modification" to refer to alterations to the test itself, although this earlier usage was not entirely consistent. We use the terminology suggested by the new *Standards*, but regardless of the terminology, it is important to note which specific alterations of standardized testing are at issue when reviewing policies or research findings.

From the perspective of measurement, the primary purpose of accommodations is to increase the validity of information about students with disabilities. In most cases, the effect of standardization of testing is to improve accuracy by eliminating irrelevant differences in testing, such as variations among schools in administrative conditions, that might distort the interpretation of scores. For example, if one school permits students 1 hour to complete a certain test, and another school allows 2 hours, the greater time allowed in the second school would bias comparisons of the two. In the case of students with disabilities, however,

standardization might itself distort interpretations, generally by biasing scores downwards because of disability-related impediments to performance.

For example, consider a student with a visual disability confronting a test of mathematics that entails substantial reading of text. Presenting that student the test in the standard format will produce a biased indication of the student's mastery of mathematics because the student's difficulty in deciphering the written page will depress his or her score. One could obtain a more accurate appraisal by changing the mode of presentation, such as using large type if the student's disability is not total, or using Braille or a taped presentation if the student's disability is so severe that it precludes reading even large type. That is, the inference about the student's mathematical proficiency would be more valid if these accommodations were provided.

The metaphor for accommodations used in a recent National Research Council report (McDonnell et al., 1997) is that of a corrective lens. A disability may distort estimates of a student's performance, perhaps causing the estimates to be too low. "Accommodations are intended to function as a corrective lens that will deflect the distorted array of observed scores back to where they ought to be—that is, back to where they provide a more valid image of the performance of individuals with disabilities" (p. 176).

Both federal statute and statute or regulations in many states require that "appropriate" accommodations be offered. But what accommodations are "appropriate?" Neither educators nor researchers have reached consensus about the appropriateness of various accommodations, and as described below, states have widely differing guidelines about their use. Several common elements do appear in many state guidelines, including requirements that assessment accommodations be consistent with the student's IEP, be used in ongoing instruction, and maintain validity and not confer unfair advantage to the student (e.g., Thurlow, House, Boys, Scott, & Ysseldyke, 2000). However, the guidelines generally do not specify what some of these criteria mean operationally. For example, the guidelines often do not specify what accommodations maintain validity and avoid conferring unfair advantage to students with specific types of disability.

Although the primary function of accommodations traditionally has been to improve the validity of scores for individuals, the recent focus of policymakers on aggregate scores—such as the performance of schools, districts, and entire

states—has led to discussion of accommodations as a means of increasing inclusion in large-scale assessments. The effects of accommodations on inclusion, however, are inevitably linked to validity. For example, assume that in one instance, providing a given accommodation has little effect on the scores of students who would have been tested without it but does result in the inclusion of additional students who would otherwise have been excluded. If the validity of inferences about the performance of these otherwise excluded students is reasonable, then the validity of conclusions about aggregate scores—such as school averages—will improve. However, allowing this accommodation might also result in educators providing it to students who otherwise would have been tested without it. This is what appears to have happened when NAEP began offering accommodations (Reese, Miller, Mazzeo, & Dossey, 1997). The validity of inferences about those students may either improve or deteriorate as a result of providing them with this accommodation, depending on the appropriateness of the accommodation for them. In practice, the mix of these effects on validity is likely to be complex, depending on the nature of the accommodations and the characteristics of the students who are allowed to use them.

One key to devising appropriate accommodations is therefore understanding both what biases—that is, what distortions in the interpretation of test scores—might be caused by a disability, and what alterations of testing might alleviate those biases without conferring inappropriate advantage. Unfortunately, we currently have limited information to answer these questions. One reason is the shortage of research. The second reason is the nature of the population with disabilities—specifically, the heterogeneity of this group and the lack of clear information about the implications for measurement of the disabilities of many students.

Inconsistent identification and classification are therefore problematic for measurement, although some educators maintain that classification is useless or even detrimental to the appropriate design of educational services. To prescribe a corrective lens, one needs to ascertain the distortion one needs to correct. If we cannot specify what a student's disability is, the task of identifying the bias caused by the disability or the types of accommodations that might lessen it becomes formidable, if not impossible. If IEP teams could consistently determine the nature of students' disabilities, the biases these disabilities cause in their assessment results, and the effect specific accommodations would have in ameliorating those distortions

without introducing new biases, then the lack of consistent classification would not pose a problem for assessment. In many cases, however, IEP teams do not have the information they need to make these judgments well. Studies have found that teachers often assign accommodations inappropriately. For example, studies have shown that teachers' decisions about accommodations for students with learning disabilities do not correspond consistently to the resulting benefits (e.g., Fuchs & Fuchs, 1999). Most forms of research that would help clarify the effects of accommodations on validity can only be carried out with meaningful classifications of students.

Disabilities Related to Measured Constructs

The example above of accommodations for visually impaired students illustrates a key point about accommodations—that is, appropriate accommodations are designed to offset impediments that are *unrelated to the construct* (i.e., the aspects of proficiency) the test is intended to measure. For example, a student's inability to read small fonts because of a visual impairment is irrelevant to her current understanding of algebra, even though it may have affected her success in studying algebra. Therefore, removing the effects of that inability by providing the test in large type or Braille will increase the validity of inferences about her mastery of algebra. Much of the policy debate about the assessment of individuals with disabilities has focused on cases in which the effects of disabilities are not relevant to the construct the test is intended to measure.

Very few identified students, however, have disabilities that have such clear implications for testing. For example, in 1996-97, only 0.05% of all students ages 6-17 were identified as having visual disabilities under IDEA Part B. Thus, such students constituted about half of one percent of the students served under IDEA. Only 0.14% of students were identified as having hearing impairments (U.S. Department of Education, 1998, Table AA13). Most identified students have disabilities that have much less clear implications for the use of accommodations. These include specific learning disabilities, emotional disturbance, and speech and language impairments.

For many students, the effects of disabilities on performance are at least in part related to the proficiencies the test is intended to measure—that is, they are construct-relevant. Moreover, in many cases, separating construct-relevant from construct-irrelevant impediments may be difficult. Students with learning disabilities, who constitute about half of all identified students with disabilities

nationwide, are a group for whom disability-related impediments are often construct-relevant. This is also generally true of students with mental retardation. A particularly clear case would be one in which a reading test is administered to a dyslexic student. The problem extends to other subject areas, however, and is exacerbated by current directions in assessment design. As a National Research Council study committee noted recently:

> By design, many performance assessments associated with standards-based reform require students to integrate a variety of knowledge and skills . . . Thus, for example, performance assessments in the area of mathematics are likely to involve reading and writing in the context of problem solving. In theory, this approach increases the probability that reading or writing disabilities, which are among the most common, will interfere with the assessment of mathematics. A similar situation exists for other assessments of other subject areas. (McDonnell et al., 1997, p. 162)

When disabilities are related to the proficiencies a test is intended to measure, designing appropriate methods of assessment and accommodation becomes extremely difficult. For example, consider a student whose learning disability impedes his reading and writing. How should such a student be assessed when a state's test involves substantial reading and writing in all subject areas, as many now do? Providing no accommodation may lead to an underestimate of the student's proficiency, but providing accommodations that offset his poor reading and writing may change the nature of the proficiencies measured by the test and produce an overestimate of the student's proficiency.

Issues Pertaining to Test Design

Increased inclusion of students with disabilities in the assessments used for most students raises a variety of issues pertaining to the design, construction, and evaluation of tests.

One issue is the possibility of item or test bias. This concern arises routinely in assessing students with backgrounds that may put them at a disadvantage in responding to a test. Techniques for addressing potential item bias are well developed, but not all assessment programs screen for possible bias involving students with disabilities. Moreover, common techniques for assessing item bias are not able to address test bias and indeed can fail even as indicators of item bias if test bias is severe. Methods for assessing bias generally require that students be matched on some criterion measure of the construct of interest. For example, if one could determine a reasonable criterion of eighth-grade mathematics proficiency, one could

match students with and without disabilities on this criterion and then determine whether students in the two groups with similar proficiency performed differently on an item or a test. Some bias-detection techniques, what Camilli and Shepard (1994) call external methods, use a criterion independent of the test in question. For example, one could use performance in college or in a job to look for bias in admissions or employment tests. Internal methods, by contrast, use part or all of the test in question as the criterion. For example, to examine possible bias on one test item, one might match students in terms of performance on the rest of the test and then look for performance differences on the item in question. As Camilli and Shepard note, internal methods are typically used to investigate item bias. These methods are suspect when there is a serious possibility that a substantial part (or the entirety) of a test may be biased for or against a particular group. And, as one of the leading figures in item-bias methodology commented, "It is impossible to seriously assess item/test 'bias' (as opposed to DIF [differential item functioning]) using performance data from a single test and nothing more" (P. Holland, personal communication, August 2002).

The increased inclusion of students with disabilities in general-education assessments therefore suggests the need for two types of investigation. First, at least over the short term, it would be prudent to increase the routine screening for item bias affecting students with disabilities. Second, with an eye to the longer term, it would be helpful to have additional research assessing the adequacy of conventional internal item-bias methods for evaluating bias affecting students with disabilities, particularly in the case of students (such as those with dyslexia) whose disabilities or accommodations might have pervasive effects across an entire test.

A second issue of design is the difficulty of the test. Many students with disabilities perform relatively poorly on tests. Using test items that are sufficiently difficult to be useful in assessing high-achieving students may lead to problems in assessing students with relatively low performance, including many students with disabilities. For example, Koretz (1997) found evidence that the statewide mathematics test in use in Kentucky at the time was too difficult for many students with disabilities. Use of test items too difficult for these students will lead to imprecise measurement and may also cause both demoralization among students and undesirable responses by teachers. This too warrants further exploration, including routine monitoring of the difficulty of new tests for students with disabilities.

The choice among assessment formats, such as multiple choice and open response, may have different implications for students with disabilities than for others. Currently, there is no agreement about the utility or fairness of different formats for students with different types of disabilities, and empirical evidence about this question is sparse (see, e.g., Koretz & Hamilton, 1999).

Review of Selected Research

Research pertaining to the testing of K-12 students with disabilities is spotty, and it leaves policymakers and educators with incomplete guidance about how best to incorporate students with disabilities into large-scale assessments. The lack of sufficient research guidance was noted by a National Research Council study 4 years ago (McDonnell et al., 1997). Although the amount of research has increased in response to the growing inclusion of students with disabilities in large-scale assessments, many key questions remain incompletely answered. For example, the literature provides substantial information about reported prevalence rates and about state policies, but it provides meager information about both the use and effects of accommodations.

Participation

Just as there are inconsistencies in identification and classification of students with disabilities, states vary markedly in terms of both policies for participation and the actual rates of participation of students with disabilities in large-scale assessments.

Participation policies. Substantial information is available on participation policies. Many descriptive studies have been conducted to determine state and national assessment program policies for including students with disabilities and the accommodations offered them (e.g., Elliott, Erickson, Thurlow, & Shriner, 2000; Erickson, Thurlow, & Ysseldyke, 1996; Mazzeo, Carlson, Voelkl, & Lutkus, 2000; McGrew, Thurlow, & Spiegel, 1993). Studies have shown that states have been developing and revising their participation policies and guidelines (Erickson et al., 1996).

Actual rates of participation. Available data on actual rates of participation are incomplete. A recent survey of state directors of special education (Thompson & Thurlow, 1999) found that only 23 states could provide the data from which participation rates could be estimated—in this case, numbers of students

participating in assessments and child count data.[2] Reasons cited for the lack of data included local variability in methods used to count students tested and the collection of data about students with disabilities by age rather than grade level.

Although states are increasing the participation of students with disabilities, Thompson and Thurlow (1999) found that the estimated rates of participation were low in some states, and it is not clear that the rates are estimated consistently. They estimated that participation rates "varied among the 23 states from less than a quarter to all students with disabilities" (p. 14). The estimate of complete inclusion for one state (Kentucky) may have been an error. Kentucky has been a leader in the push to increase inclusion, but complete inclusion is implausible. Some students will always be excluded, even if unintentionally, for a variety of reasons, including illness, truancy, parental unwillingness to have their children tested, and so on. Other studies (e.g., Koretz, 1997) have estimated high but not nearly complete inclusion in Kentucky's assessment. One possible explanation for the reporting of complete inclusion is that all students are counted in Kentucky's accountability system, and a student who fails to take the assessments is counted as a zero. In that sense, all students are included in the accountability system, even if they are not actually administered the assessment. This implausibly high estimate underscores the need for a standardized method for estimating participation rates.

State Policies About Accommodations

The availability and use of accommodations in state assessments is determined both by state guidelines for the use of accommodations and by local decisions, particularly the decisions of each student's IEP team.

Surveys of state policies pertaining to accommodations conducted over the past decade (National Center on Economic Outcomes [NCEO], 1993; Thurlow et al., 2000; Thurlow, Seyfarth, Scott, & Ysseldyke, 1997) show them to be both inconsistent and in a state of rapid flux. The 1993 NCEO survey identified four types of accommodations used by many state and federal testing agencies, but states varied in terms of which accommodations they allowed and forbad, and no single accommodation was permitted in every state. The later surveys showed rapid changes in policy and continuing inconsistencies among states, but some consistent themes did become apparent. Thurlow et al. (1997) analyzed policy changes to 1997,

[2]Child counts are usually obtained in the fall, and test data are usually obtained in the spring, so estimates of participation based on these data are only approximate.

the year in which IDEA was last reauthorized. They found rapid changes in policies pertaining to both participation of students with disabilities and accommodations. A growing number of states had begun offering either partial participation in the general assessment or an alternate assessment for certain students with disabilities. States were increasingly focused on the issue of accommodations; Thurlow et al. (1997) found that nearly every state with published assessment guidelines included guidelines for accommodations, which was not true at the time of the 1993 NCEO survey. They also noted a decrease in attention to disability categories in state guidelines. Among the accommodations that showed the most inconsistency among states (in the sense that some states explicitly approved it, whereas others explicitly forbad it) was reading a test aloud. Policies permitting the use of scribes were common.

Thurlow et al. (2000) reexamined state policies in late 1999, 2 years after the reauthorization of IDEA. Although IDEA now requires that IEP teams make decisions about accommodations, states continue to provide guidelines for how they should do so. Thurlow et al. (2000) noted a continuing decrease in attention to disability categories in many aspects of the guidelines, but also noted a return to consideration of category in some respects, particularly in making decisions about participation in alternate assessments. They found an increase in the specification of criteria for local decision making about the use of accommodations. The two most common of these criteria pertained to the role of the IEP in making decisions and the criterion that assessment accommodations must be used in instruction as well. State guidelines have increasingly distinguished between "standard" accommodations (those specified in the guidelines) and "nonstandard" accommodations, and some states specified that scores resulting from nonstandard accommodations would not be aggregated into reported statistics.

Most states explicitly offered Braille or large-print editions of the tests (Thurlow et al., 1997, 2000). Other frequently offered accommodations including proctor or scribe (75% in 1997 and 94% in 2000) and extended time (33% in 1997 and 77% in 2000). The most commonly offered setting accommodation was allowing students to take the test individually or in a small group; a substantial majority of states explicitly permitted this.

Recent federal legislation is likely to contribute to continuing changes in policies pertaining to the use of accommodations and is also likely to make those policies more salient to local educators. For example, Thurlow (2001) reported that,

as a result of this legislation, all states now have written guidelines pertaining to the use of accommodations, and state education department staff are increasingly aware of them. Policies about accommodations are likely to continue to evolve in the face of continuing policy efforts and accumulating experience.

Actual Use of Accommodations in Large-Scale Assessments

State policies shape the use of accommodations but do not indicate the frequency with which accommodations are actually used in practice. Actual use depends on local decisions, as mandated by federal statute.

Studies of the use of accommodations do not paint an entirely consistent picture of accommodation use. It would be surprising if the findings were consistent. One would expect variations between states because of differences in guidelines, the characteristics of the assessments used, and the size and characteristics of the identified population. Moreover, some information on the use of accommodations comes from data collected in operational assessment systems, and different systems collect different information about accommodations. Finally, one would expect variation among localities within states, as well as between them, because of differences in the characteristics of the students identified and of the educators serving them.

Nonetheless, several accommodations were found in numerous studies to be particularly common. Several studies have found that the most often used accommodations are extended time (Elliott, Kratochwill, & Schulte, 1998; Marquart, 2000; Mazzeo et al., 2000) and reading items (Elliott & McKevitt, 2000) or directions (Mazzeo et al., 2000) aloud. In a recent study, McKevitt (2000) found that, on performance assessments

> [t]he most frequent accommodations used were verbal encouragement of effort (used with 60 students), read directions to student (60 students), simplify language in directions (55 students), reread subtask directions (54 students), have student restate directions to the teacher in his/her own words (49 students), read test questions and content to student (46 students), and restate questions with more appropriate vocabulary (46 students). Extra time was necessary for 31 students with disabilities, but offered for almost all. (p. 4)

Because current policy focuses on increasing the inclusion of students with disabilities in the state assessments administered to general-education students, it is particularly important to look at the use of accommodations under those

circumstances. Two CRESST studies explored this issue in Kentucky, which was one of the first states to succeed in including the large majority of its students with disabilities in its regular large-scale assessment. These studies showed that the use of accommodations was extensive but varied by grade, decreasing as students progressed through school. Of the 4th-grade students with disabilities included in the regular assessment, 81% received at least one accommodation, and 66% received two or more. These percentages dropped to 61% and 41%, respectively, in 11th grade (Koretz, 1997, Table 5; see also Trimble, 1998). These percentages are especially high given that extended time, the accommodation found to be most common in some other studies, was not counted as an accommodation in Kentucky.

CRESST's studies showed that in Kentucky, as in other studies, reading either directions or items—called "oral presentation" in Kentucky—was a very common accommodation, particularly for elementary school students (Table 2). However, these studies also showed that, in practice, even accommodations intended for occasional use may be offered widely. For example, Kentucky's guidelines at the time presented paraphrasing as an intrusive technique, and they admonished educators to use the least intrusive accommodations possible. Nonetheless, almost three fourths of the students with disabilities in the fourth grade and about half of the students with disabilities in the secondary grades were provided with paraphrasing (Table 2).

A recent survey of states (Thompson & Thurlow, 1999) provided estimates of the percentages of students with disabilities assessed with accommodations in 12 states. These estimates ranged from 8% (Kansas, Grade 10) to 82% (Kentucky, Grade 4, and Indiana, Grade 10; see Thompson & Thurlow, 1999, Table 7). These estimates, however, are not comparable to each other. First, the reported percentages of students with disabilities assessed (with or without accommodations) varied from 32% (Pennsylvania, Grade 10) to over 90% (Kentucky, Rhode Island, Massachusetts, and Maryland; see Thompson & Thurlow, 1999, Table 6). In some instances, lower rates of accommodation may simply reflect lower rates of participation by students with relatively severe disabilities. Second, the differences in rates may reflect differences in which accommodations were reported. For example, Thompson and Thurlow (1999) found that some states reported only modifications of the tests employed, which would exclude the particularly common accommodations.

Table 2

Percentage of Kentucky Students With Disabilities Receiving Assessment Accommodations, by Grade, 1995

Accommodation	Grade 4	Grade 8	Grade 11
None	19	33	39
Oral presentation	72	56	45
Paraphrasing	49	48	47
Dictation	50	14	5
Cueing	10	12	10
Technological aid	3	5	5
Interpreter	2	3	4
Other	8	5	6

Note. Individual students may receive multiple accommodations. From D. Koretz, *Assessment of students with disabilities in Kentucky* (CSE Tech. Rep. No. 431), University of California, Los Angeles, Center for Research on Evaluation, Standards, and Student Testing, 1997, p. 13 (Table 6).

Effects of Accommodations on Participation

Although one rationale for accommodations is to improve participation by removing irrelevant impediments to assessment, actual data on the effects of accommodations on participation are sparse. One reason for the paucity of information is that there have been few instances in which a decision was made to begin offering accommodations where none at all had been offered previously.

One source of information is NAEP, which first began offering accommodations in 1996. The initial provision of accommodations in NAEP was carefully planned to permit comparisons of similar samples, one with accommodations offered and one without. When accommodations were first introduced (in the 1996 mathematics assessment), their provision had two effects:

- The provision of accommodations and adaptations clearly increased participation rates for students with disabilities and LEP students at grades 4 and 8. When accommodation or adaptations were available, more than 70% of both of these groups were assessed at each of these two grades. These numbers are substantially higher than the program has achieved in past assessments that did not offer accommodations and adaptations. On the other hand, providing accommodations at grade 12 had little effect on participation.

- A portion of the population of students with disabilities was assessed with accommodations or adaptations when these were available but was

> assessed under standard conditions when these special administration procedures were not available. A similar pattern of results was not evident among LEP students. (Reese et al., 1997, p. 65).

However, somewhat different patterns were found in the 2000 mathematics assessment, which again included two comparable samples, of which only one was offered accommodations. Again, somewhat more students with disabilities were assessed in the sample that permitted accommodations. However, in 2000, the provision of accommodations did not appear to substantially change the percentage of students with disabilities assessed under standard conditions. Only additional studies will resolve this inconsistency and show how often, and under what circumstance, the provision of accommodations changes the proportion of students with disabilities included but assessed without accommodations.

Effects of Accommodations on the Quality of Performance Information

As noted above, the primary purpose of accommodations is to increase the validity of information about the performance of students with disabilities. Ideally, by removing irrelevant impediments to performance on a test, the use of accommodations can provide a more accurate indication of the knowledge and skills of students with disabilities.

Unfortunately, research about the effects of accommodations on validity remains very limited, and only a modest share of the relevant research explores the effects of accommodations in K-12 assessments. (Other research has focused on tests used for admission to college, graduate school, and professional schools.) However, the research literature is rapidly growing.

The available research is clouded by the lack of an unambiguous criterion for the performance of students with disabilities. The purpose of accommodations is to make the information provided by a given test more similar to the information from a measure that is not biased by a student's disability. However, in many circumstances, it is not practical to identify another measure that is clearly unbiased or less biased than the test in question. For example, say that the issue is which of two accommodations provides the more valid information from a given test for a given type of student, and validation evidence will come from a comparison with another test. The same ambiguity is likely to arise with the second test: Does it produce more valid information with accommodations or without? Much the same issue arises in comparing test scores to other variables, such as teachers' grades. Are

the grades a teacher assigns to a given student with disabilities less biased, equally biased, or more biased than scores on a test administered under standard or accommodated conditions?

An exception to this general lack of a clear criterion is tests that are used to predict later performance, such as college admissions tests. In these cases, future performance is the criterion, and the accuracy of predictions with and without accommodations can be compared. However, most K-12 tests are not explicitly designed to support predictions. Phillips (1994) suggested using internal evidence instead, specifically that the impact of an accommodation on validity can be gauged in part by whether it benefits students with disabilities without similarly benefiting all students. Fuchs and Fuchs (1999) have developed a diagnostic testing procedure based on this premise, in which the effects of an accommodation for a specific student are compared to normative data on the effects for students without disabilities. Elliott and McKevitt (2000), however, argued that defining the validity of an accommodation in such a limited manner may inadvertently restrict the availability of accommodations that may actually improve validity for certain students. Other responses to the lack of a criterion have also relied on internal evidence. Koretz (1997) examined whether differences in scores between accommodated and other students appeared plausible but cautioned that these judgments are warranted only for the relatively few cases in which one has a clear basis for expecting a given distribution of scores. Other studies have examined the psychometric properties of scores of students with and without accommodations (e.g., Koretz, 1997; Koretz & Hamilton, 1999; Willingham et al., 1988).

Studies of accommodations in college admissions and graduate admissions tests. A considerable amount of research has explored the assessment of students with disabilities with the SAT and Graduate Record Examination (GRE) (Willingham et al., 1988). This research examined the performance of students with disabilities tested under both standard and accommodated conditions. These studies considered numerous disability categories separately, specifically for students with hearing impairments, visual impairments, physical disabilities, and learning disabilities. They examined issues pertaining to the identification of students with disabilities, the use of accommodations, the psychometric properties of scores from standard and nonstandard administrations, and the predictive value of scores obtained with accommodations.

Internal psychometric evidence—that is, evidence from performance on the tests themselves—was generally positive. For example, scores on the tests showed similar factor structures for students with and without disabilities. There was also little sign of differential item functioning (that is, anomalous performance on individual items) for students with certain disabilities. These findings suggest that the test was to a considerable degree measuring the same things for both groups. The reliability of scores was also similar for the two groups (Bennett, Rock, Kaplan, & Jirele, 1988; Rock, Bennett, Kaplan, & Jirele, 1988).

However, the primary function of these tests is to predict later performance in college or graduate school, and in that respect, the findings provided more grounds for concern. In general, scores for students with disabilities who were assessed with accommodations did not predict later performance (grade-point average [GPA] in college or graduate school) as accurately as did scores for students with no disabilities. This effect was small overall but sizable for some groups. In most instances, scores for students provided with accommodations overpredicted later GPA. That is, on average, students provided accommodations received lower grades subsequently than would have been predicted on the basis of their test scores. This was particularly true of high-performing students. The exception was students with hearing disabilities, who received higher-than-predicted freshman grades regardless of accommodations (Braun, Ragosta, & Kaplan, 1988). It is important to note, however, that if some colleges and universities at the time of these studies were not providing appropriate accommodations to some students with disabilities, the freshman GPA of those students may have been depressed as a result. To the extent that this was true, the overprediction of GPA could have reflected bias in GPA rather than in test scores. Therefore, it would be valuable to replicate the analysis of predictive accuracy using more recent data.

These studies also explored the need for extra time by comparing the rates at which students with disabilities (but without accommodations) and students without disabilities completed the test. Students with disabilities did not have a lower rate of completion, and students with certain disabilities actually had higher rates of completion (Bennett et al., 1988).

Given the nature of the samples and the tests in these studies, it would be risky to generalize the findings to K-12 assessments. They do, however, raise two issues that must be investigated for these assessments. First, they suggest that accommodations may be used excessively in some cases. Second, they suggest that,

in some instances, accommodations may undermine rather than strengthen validity, perhaps because of the lack of standardization of accommodations themselves (Braun et al., 1988).

Extended time. One of the most researched accommodations has been the use of extended time. The conclusions of this research are mixed, presumably in part because tests vary in their speededness (that is, the extent to which performance depends on speed of work). Some studies suggest little difference between the performance of students with or without disabilities under extended time conditions (Fuchs, Fuchs, Eaton, Hamlette, & Karns, 2000; Halla, 1988; Linder, 1989; Marquart, 2000; Munger & Lloyd, 1991). Others have found that students both with and without disabilities performed better with more or unlimited amounts of time (Alster, 1997; Centra, 1986; Gallina, 1989; Hill, 1984; Huesman & Frisbie, 2000; Linder, 1989; Montani, 1995; Ofiesh, 1997; Perlman, Borger, Collins, Elenbogen, & Wood, 1996; Runyon, 1991a, 1991b; Weaver, 1993). Research has not yet resolved these inconsistencies. Moreover, in the cases where extended time does have an impact, it is not yet generally known how much extra time is appropriate for students with various disabilities, although some limited information is available. Packer (1987) looked at variations in required testing time for students receiving different accommodations while taking the Scholastic Aptitude Test (SAT) from years 1979-1983. She found visually impaired students using a Braille form took longer than those using cassette versions; physically and learning disabled students using large print or regular type and learning disabled students using cassette versions took similar amounts of time, less than the visually impaired students, and hearing impaired students taking the regular version needed the least amount of time.

Reading aloud. Reading test items aloud has become an increasingly debated and researched accommodation. Reading aloud on math or science assessments is widely accepted. However, reading aloud on a reading test is, predictably, highly controversial. Some believe the read-aloud accommodation on a reading test changes what is being measured to listening comprehension (Meloy, Frisbie, & Deville, 2000; Phillips, 1994). Others maintain that a read-aloud truly levels the playing field for students with reading difficulties (Elliott, Ysseldyke, Thurlow, & Erickson, 1998; Harker & Feldt, 1993). The answer to this disagreement must be grounded in the purpose(s) of the assessment and, in particular, the specific inferences the test is intended to support. For example, inferences about language

comprehension may be warranted under read-aloud conditions, but inferences about the ability to extract information from written text would not be.

Most studies conducted on the effects of reading tests aloud have examined performance on mathematics tests. Some have found students with disabilities do better when the tests are read to them (Bennett, Rock, & Kaplan, 1987; Espin & Sindelar, 1988; Koretz, 1997; Meloy et al., 2000; Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998; Weston, 1999) In contrast, Tindal, Almond, Heath, and Tedesco (1998) and Trimble (1998) found students with disabilities showed no performance increase when the test was read to them. Harker and Feldt (1993) found students without disabilities classified as weak readers did better with a cassette-recorded version of a reading test, but strong readers did better without the recording.

Concurrent accommodations. Most students with disabilities receive more than one accommodation. In some instances, the nature of a disability may clearly suggest the need for multiple accommodations, in some cases because one accommodation may necessitate others. For example, a hearing impaired student requiring signed instructions may also require extended time (because of the pace of signing) and possibly a separate testing room, which may even entail an individual administration. In other cases, however, the reasons for multiple accommodations are unclear, at least for those unfamiliar with the individual students and their IEPs.

The extent of multiple accommodations is illustrated by a study of the Kentucky assessment, which was one of the first general-education assessments to include the large majority of students with disabilities. Teachers proctoring the Kentucky assessment were asked to indicate which of seven categories of accommodations (six specific accommodations and a seventh category for "other") were provided to each student. Koretz (1997) found 59 different combinations of two or more accommodations among fourth-grade students with disabilities. Students were two to four times as likely to receive multiple accommodations as single accommodations, depending on grade (Koretz, 1997; see Table 3).

A number of studies have examined the apparent effects of multiple accommodations but have yielded inconsistent findings. Some studies found that accommodations help students with disabilities improve test performance (Perez, 1980; Tachibana, 1986); others found the opposite, or even sundry results. For example, in studies that included various concurrently administered

Table 3

Percentage of Kentucky Students With Disabilities Receiving Assessment Accommodations, by Grade, 1995

Accommodation	Grade 4	Grade 8	Grade 11
No accommodations	19	33	39
One accommodation	15	22	20
Multiple accommodations	66	45	41

Note. From D. Koretz, *Assessment of students with disabilities in Kentucky* (CSE Tech. Rep. No. 431), University of California, Los Angeles, Center for Research on Evaluation, Standards, and Student Testing, 1997, p. 12 (Table 5).

accommodations, such as read aloud, extended time, pacing,scribes, large print, and/or cueing, students showed score gains (Burk, 1998), score gains that were questionable (Koretz, 1997), no gains in scores (Helwig, Rozek-Tedesco, Heath, Tindal, & Almond, 1999; Mick, 1989; Schulte, Elliott, & Kratochwill, 2000; Tindal, Glasgow, Helwig, Hollenbeck, & Heath, 1998), score losses (Trimble, 1998), or differential score gains (Fuchs et al., 2000).

The inconsistency of the findings of studies of multiple accommodations are unsurprising, given the variation among tests, disabilities, and the combinations of accommodations offered. In addition, the use of multiple accommodations further complicates efforts to ascertain the effects of accommodations. When so many different combinations of accommodations are provided, the number of students receiving any given combination is typically very small, often far too small to permit reliable analysis of any single group. Statistical methods for disentangling the effects of multiple accommodations are weak under these circumstances (see Koretz, 1997).

Implications for Research

In mandating the inclusion of students with disabilities in the general-education assessments used in standards-based reform, the nation moved into uncharted territory. Ideally, at the outset of a reform of this magnitude, policymakers and educators would have both solid information about the problem at which the reform is targeted and research-based evidence suggesting the range of likely effects, both positive and negative, of various policy responses. The former would include, for example, firm data on the numbers of students with various disabilities, the ways in which those students were assessed, and the distribution of achievement and attainment among students with disabilities. Research-based

evidence that would help predict the effects of new policies would include information on appropriate and effective ways of assessing students with disabilities, information on the effects of accommodations, and evidence suggesting the effects of including various types of students with disabilities in statewide general-education testing and accountability programs. In this case, as the review above shows, much of the potentially valuable information was weak or entirely lacking. Certainly, policymakers and educators had information addressing some of these questions. For example, there was sufficient evidence about the distressingly low rates of educational attainment of students with disabilities (such as their high-school completion rates) to make clear that the education of many students with disabilities was inadequate. Overall, however, the available systematic evidence was limited and often weak.

Current policy is therefore based partly on unsubstantiated assumptions—for example, assumptions that any statewide assessment will be appropriate for the great majority of students with disabilities and that only a very small percentage will need an alternate assessment; that accommodations specified by IEP teams under inconsistent guidance from states will result in valid information about student performance; and that the result of inclusion in these assessments will be improved outcomes for students with disabilities. Although the reforms of the past decade and half have spurred a sizable increase in research, the available research-based information remains seriously inadequate. If we are to monitor these reforms effectively and decide where midcourse corrections are needed, we will need to generate a great deal of additional information of several types.

To start, we need better descriptive information about the target populations, the contexts in which they are schooled, and the ways in which they are assessed. We cannot effectively evaluate the new assessment policies if we cannot identify the target groups clearly, and the current dramatically inconsistent data on prevalence suggest that we are failing to do so. We need to know how students are being assessed—for example, what proportions of students of various types are in fact included in the general-education assessments, and what accommodations they are provided. This information should be collected in a manner that allows one to differentiate among types of students, schooling contexts, and types of assessment and accountability systems.

To obtain an adequate and comparative perspective on these questions, we need greater standardization of definitions and data collection methods. An analogy

is the need for consistent information about dropout rates. Historically, states, local districts, and other organizations calculated dropout rates using a wide variety of disparate and often misleading methods. In recent years, the policy community has come to recognize that if they are to act effectively on information about dropout rates, they need rates that are calculated using standard and well-understood methods and that therefore permit meaningful comparisons. The current reforms in the assessment of students with disabilities pose similar issues. As the examples above indicate, we cannot draw fully adequate inferences about even some simple questions, such as participation rates and patterns in the use of accommodations, if there is no consistency across jurisdictions in the definition of terms or the methods used to obtain data.

Second, there is a pressing need for additional research on the assessment of students with disabilities. The needed research is diverse; some can be accommodated through routine data collection in the course of developing or administering the assessments, while other studies will require much more demanding designs. We need further investigation of possible item bias, test bias, and excessive difficulty affecting students with disabilities taking statewide general-education assessments. One of the most urgent questions is the overarching question of evidential validity—that is, the degree to which the results of assessments of students with disabilities support desired inferences about their academic performance. Operational assessment data can provide only limited information about validity because of the lack of control over the assignment of accommodations and the lack of information about the characteristics of the students provided different accommodations, such as their disabilities and the reasons for their IEP teams' decisions about accommodations. Accordingly, firm answers about the validity of scores when accommodations are used may require experimental designs.

Research on the assessment of students with disabilities also needs to become more context-sensitive and comparative. There is no reason to expect the functioning of assessment systems to be similar across a wide variety of assessment systems or throughout the extremely heterogeneous group of students classified as disabled. As noted above, the findings of extant research are in some instances highly inconsistent. It is very difficult to make sense of these variations in findings when studies differ simultaneously in terms of the characteristics of the populations investigated, the types of assessments employed, the use of accommodations, and

the methods used to define and collect information about key variables (such as the use of accommodations). If research evolves toward a more explicitly comparative framework, with greater emphasis on describing the factors that could explain variations in findings and greater similarity of design across studies, comparisons among studies will gradually yield more useful information.

Finally, there is an urgent need to evaluate the diverse effects of these policies on the learning, attainment, and well-being of students with disabilities. Strong voices can be heard arguing about both benefits and harms accruing to students with disabilities as a result of the new policies, and anecdotal evidence can be found to buttress both views. Particularly in the light of the weak evidentiary basis undergirding some aspects of the reforms, the importance of rigorous and broadly focused evaluation cannot be overstated. To be realistic and useful, this research must be sensitive to differences among types of students, educational contexts, and assessment systems, as what works well with one type of student or in one type of context may work very poorly in another.

References

Alster, E. H. (1997). The effects of extended time on algebra test scores for college students with and without learning disabilities. *Journal of Learning Disabilities, 30,* 222-227.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological tests.* Washington, DC: American Educational Research Association.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement, 24,* 44-55.

Bennett, R. E., Rock, D. A., Kaplan, B. A., & Jirele, T. (1988). Psychometric characteristics. In W. W. Willingham, M. Ragosta, R. E. Bennett, H. Braun, D. A. Rock, & D. E. Powers (Eds.), *Testing handicapped people* (pp. 83-98). Boston: Allyn and Bacon.

Braun, H., Ragosta, M., & Kaplan, B. (1988). Predictive validity. In W. W. Willingham, M. Ragosta, R. E. Bennett, H. Braun, D. A. Rock, & D. E. Powers (Eds.), *Testing handicapped people* (pp. 109-132). Boston: Allyn and Bacon.

Burk, M. (1998, October). *Computerized test accommodations: A new approach for inclusion and success for students with disabilities.* Paper presented at the Office of Special Education Program Cross Project Meeting "Technology and the Education of Children with Disabilities: Steppingstones to the 21st Century."

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

Centra, J. A., (1986). Handicapped student performance on the Scholastic Aptitude Test. *Journal of Learning Disabilities, 19,* 324-327.

Clarizio, H. F., & Phillips, S. E. (1992). A comparison of severe discrepancy formulae: Implications for policy consultation. *Journal of Educational and Psychological Consultation, 3,* 55-68.

Elliott, J. L., Erickson, R. N., Thurlow, M. L., & Shriner, J. G. (2000). *State-level accountability for the performance of students with disabilities: Five years of change?* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Elliott, S. N., Kratochwill, T. R., & Schulte, A. G. (1998). The assessment accommodations checklist: Who, what, where, when, why, and how? *Teaching Exceptional Children, 31*(2), 10-14.

Elliott, S. N., & McKevitt, B. C. (2000, April). *Testing accommodations decisions: Legal and technical issues challenging educators or "good" test scores are hard to come by.*

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Elliott, J. L., Ysseldyke, J., Thurlow, M., & Erickson, R. (1998). What about assessment and accountability? Practical implications for educators. *Teaching Exceptional Children, 31*(1), 20-27.

Erickson, R. N., & Thurlow, M. L. (1996). *State special education outcomes 1995.* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Erickson, R .N., Thurlow, M. L., & Thor, K. (1995). *1994 State special education outcomes.* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Erickson, R. N., Thurlow, M. L., & Ysseldyke, J. E. (1996). Drifting denominators: Issues in determining participation rates for students with disabilities in statewide assessment programs. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Espin, C. A., & Sindelar, P. T. (1988). Auditory feedback and writing: Learning disabled and nondisabled students. *Exceptional Children, 55,* 45-51.

Fuchs, L. S., & Fuchs, D. (1999, November). Fair and unfair testing accommodations. *The School Administrator Web Edition.* Retrieved January 16, 2002 from http://www.aasa.org/publications/sa/1999_11/fuchs.htm

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C., & Karns, K. (2000). Supplementing teacher judgments of test accommodations with objective data sources. *School Psychology Review, 29,* 65-85.

Gallina, N. B. (1989). Tourette's syndrome children: Significant achievement and social behavior variables (Tourette's syndrome, attention deficit hyperactivity disorder) (Doctoral dissertation, City University of New York, 1989). *Dissertation Abstracts International, 50,* 0046.

Halla, J. W. (1988). A psychological study of psychometric differences in Graduate Record Examinations General Test scores between learning disabled and non-learning disabled adults (Doctoral dissertation, Texas Tech University, 1988). *Dissertation Abstracts International, 49,* 0230.

Harker, J. K., & Feldt, L. S. (1993). A comparison of achievement test performance of non-disabled students under silent reading and reading plus listening modes of administration. *Applied Measurement in Education, 6,* 307-320.

Helwig, R., Rozek-Tedesco, M., Heath, B., Tindal, G., & Almond, P. (1999). Reading as an access to mathematics problem solving on multiple choice tests for sixth-grade students. *Journal of Educational Research, 93,* 113-125.

Hill, G. A. (1984). Learning disabled college students: The assessment of academic aptitude (Doctoral dissertation, Texas Tech University, 1984). *Dissertation Abstracts International, 46,* 0230.

Huesman, R. L., & Frisbie, D. A. (2000, April). *The validity of ITBS Reading comprehension test scores for learning disabled and non learning disabled students under extended-time conditions.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Individuals with Disabilities Education Act Amendments of 1997, Pub. L. No. 105-17, 37 Stat. 111 (1997).

Koretz, D. (1997). *Assessment of students with disabilities in Kentucky* (CSE Tech. Rep. No. 431). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., & Hamilton, L. (1999, January). *Assessing students with disabilities in Kentucky: The effects of accommodations, format, and subject* (CSE Tech. Rep. No. 498). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Linder, E. A. (1989). *Learning disabled college students: A psychological assessment of scholastic aptitude.* Unpublished doctoral dissertation, Texas Tech University, Lubbock.

Marquart, A. M. (2000, April). *The use of time as an accommodation on a standardized mathematics tests: An investigation of effects on scores and perceived consequences for students of various skill levels.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Mazzeo, J., Carlson, J. E., Voelkl, K. E., & Lutkus, A. D. (2000). *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities* (NCES 2000-473). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

McDonnell, L., McLaughlin, M., & Morison, P. (Eds.). (1997). *Educating one and all: Students with disabilities and standards-based reform.* Washington, DC: National Academy Press.

McGrew, K. S., Thurlow, M. L., Shriner, J. G., & Spiegel, A. N. (1992). *Inclusion of students with disabilities in national and state data collection programs* (Tech. Rep. 2). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

McGrew, K. S., Thurlow, M. L., & Spiegel, A. N. (1993). An investigation of the exclusion of students with disabilities in national data collection programs. *Educational Evaluation and Policy Analysis, 15,* 339-352.

McKevitt, B. C. (2000, June). *The use and effects of testing accommodations on math and science performance assessments.* Paper presented at the Council of Chief State School Officers (CCSSO) Annual National Conference on Large-Scale Assessment, Snowbird, UT.

Meloy, L. L., Frisbie, D., & Deville, C. (2000, April). *The effect of a reading accommodation on standardized test scores of learning disabled and non-learning disabled students.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement,* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Mick, L. B. (1989). Measurement effects of modifications in minimum competency test formats for exceptional students. *Measurement and Evaluation in Counseling and Development, 22,* 31-36.

Montani, T. O. (1995). Calculation skills of third-grade children with mathematics and reading difficulties (learning disabilities) (Doctoral dissertation, Rutgers the State University of New Jersey, 1995). *Dissertation Abstracts International, 56,* 0910.

Munger, G. F., & Lloyd, B. H. (1991). Effect of speededness on test performance of handicapped and nonhandicapped examinees. *Journal of Educational Research, 85* , 53-57.

National Center on Educational Outcomes. (1993, March). *Testing accommodations for students with disabilities: A review of the literature* (Synthesis Rep. 4). St. Cloud State University & National Association of State Directors of Special Education. Minneapolis: University of Minnesota, National Center on Educational Outcomes. (ERIC Document Reproduction Service No. ED 358 656)

Ofiesh, N. S. (1997). Using processing speed tests to predict the benefit of extended test time for university students with learning disabilities (Doctoral dissertation, The Pennsylvania State University, 1997). *Dissertation Abstracts International, 58,* 0176.

Packer, J. (1987). *SAT testing time for students with disabilities.* Princeton, NJ: Educational Testing Service.

Perlman, C. L., Borger, J., Collins, C. B., Elenbogen, J. C., & Wood, J. (1996, April). *The effect of extended time limits on learning disabled students' scores on standardized reading tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Perez, J. V. (1980). Procedural adaptations and format modifications in minimum competency testing of learning disabled students: A clinical investigation (Doctoral dissertation, University of South Florida, 1980). *Dissertation Abstracts International, 41,* 0206.

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7,* 93-120.

Reese, C. M., Miller, K. E. Mazzeo, J., & Dossey, J. A. (1997). *NAEP 1996 mathematics report card for the nation and the states: Findings from the National Assessment of Educational Progress* (NCES No. 97488). Washington, DC: National Center for Education Statistics.

Rock, D. A., Bennett, R. E., Kaplan, B. A., & Jirele, T. (1988). In W. W. Willingham, M. Ragosta, R. E. Bennett, H. Braun, D. A. Rock, & D. E. Powers (Eds.),*Testing handicapped people* (pp. 99-108). Boston: Allyn and Bacon.

Runyon, M. K. (1991a). The effect of extra time on reading comprehension scores for university students with and without learning disabilities. *Journal of Learning Disabilities, 24,* 104-108.

Runyon, M. K. (1991b). *Reading comprehension performance of learning-disabled and non-learning disabled college and university students under timed and untimed conditions.* Unpublished doctoral dissertation, University of California, Berkeley.

Schulte, A. A. G., Elliott, S. N., & Kratochwill, T. R. (2000, June). Experimental analysis of the effects of testing accommodations on students' standardized mathematics test scores. Paper presented at the Council of Chief State School Officers (CCSSO) Annual National Conference on Large-Scale Assessment, Snowbird, UT.

Shepard, L. A. (1989). Identification of mild handicaps. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 545-572). New York: American Council on Education/Macmillan.

Shriner, J. G., & Thurlow, M. L. (1992). *State special education outcomes 1991.* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Tachibana, K. K. (1986). Standardized testing modifications for learning disabled college students in Florida (modality) (Doctoral dissertation, University of Miami, 1986). *Dissertation Abstracts International, 47,* 0125.

Thompson, S., & Thurlow, M. (1999, December). *1999 State special education outcomes: A report on state activities at the end of the century.* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. (2001). *Use of accommodations in state assessments: What databases tell us about differential levels of use and how to document the use of accommodations* (Tech. Rep. 30). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., House, A., Boys, C., Scott, D., & Ysseldyke, J. (2000). *State participation and accommodation policies for students with disabilities: 1999 update* (Synthesis

Rep. 33). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., Seyfarth, A. L., Scott, D., & Ysseldyke, J. (1997). *State assessment policies on participation and accommodations for students with disabilities: 1997 Update* (Synthesis Rep. 29). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Tindal, G., Almond, P., Heath, B., & Tedesco, M. (1998). *Single subject research using audio cassette read aloud in math.* Manuscript submitted for publication, University of Oregon.

Tindal, G., Glasgow, A., Helwig, B., Hollenbeck, K., & Heath, B. (1998). *Accommodations in large scale tests for students with disabilities: An investigation of reading math tests using video technology.* Unpublished manuscript with Council of Chief State School Officers, Washington, DC.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64,* 439-450.

Trimble, S. (1998). *Performance trends and use of accommodations on a statewide assessment* (Maryland/Kentucky State Assessment Series Rep. No. 3). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

U.S. Department of Education. (1998). *To assure the free appropriate public education of all children with disabilities: 20th Annual report to Congress on the implementation of the Individuals With Disabilities Education Act.* Washington, DC: Author.

U.S. Department of Education. (2001). *To assure the free appropriate public education of all children with disabilities: 23rd Annual report to Congress on the implementation of the Individuals With Disabilities Education Act.* Washington, DC: Author. Retrieved August 30, 2002, from http://www.ed.gov/offices/OSERS/OSEP/Products/OSEP2001An1Rpt/Appendix_A_Pt1.pdf

Weaver, S. M. (1993). *The validity of the use of extended and untimed testing for postsecondary students with learning disabilities (extended testing).* Unpublished doctoral dissertation, University of Toronto.

Weston, T. (1999, April). *The validity of oral presentation in testing.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A. & Powers, D. E. (Eds.). (1988). *Testing handicapped people.* Boston: Allyn and Bacon.

Ysseldyke, J. E., & Algozzine, B. (1982). Bias among professionals who erroneously declare students eligible for special services. *Journal of Experimental Education, 50,* 223-228.

BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").